コーパスを利用した日本語語形DBの構築　ー動詞語形を例としてー

川　田　　亮　一
熊本学園大学商学部・経営学科
kawada@kumagaku.ac.jp

１　はじめに

ここ数年多くの電子化テキストやコーパスが利用可能になってきている。これらは言語処理システム・手法の検証や言語知識の自動抽出などのデータとして用いられることが多いが、言語研究の立場からも有用である。筆者は広く言語研究の立場から、電子化コーパスを利用した日本語語形データベースの構築を目指しており、その一部として動詞語形データベースを作成した。本稿では、この動詞語形データベース作成を例として、電子化された一次言語データから二次言語データを作成する方法及び問題点を報告する。

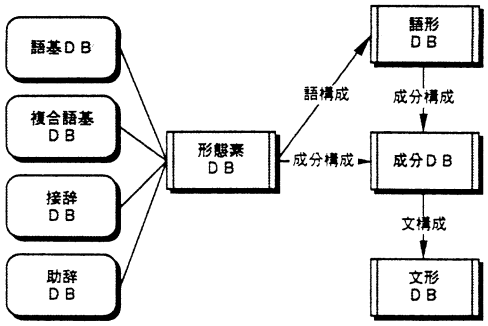
２　日本語語形DB

２．１　語形DBの位置づけ

本稿で「語形」と呼ぶのは、いわゆる「文節形」のことである。筆者は形態素（いわゆる語）が語（いわゆる文節）を構成する仕組み＝語構成論を形態論、できあがった語が文の成分を構成する仕組み＝成分論、成分が文を構成する仕組み＝構文論とする文法観[1]に立っているが、術語の紛らわしさを避けるため、本稿ではいわゆる文節形を「語形」と呼ぶことにする。

さてこの立場に立つと、主に文法との関係から図１ような言語データベースを想定することが出来る。語形データベースとは形態素が構成した語形の一覧表ということになる。同種のものには、国立国語研究所の「文節形度数表」[2]がある。

図１．語形ＤＢの位置づけ



２．２　動詞語形DBの位置づけ

語形データベースは、形態素が構成した語形の一覧であるから、以下のように形態素（語基）の種類別に作成するのが妥当である。

体言語基　→　名詞語形ＤＢ
用言語基　→　動詞語形ＤＢ
形容語基　→　形容詞語形ＤＢ
…
感動詞ＤＢ
接続詞ＤＢ
連体詞ＤＢ

ただし、感動詞～連体詞は既に語形としての資格を持っているために見出し語そのものの一覧となるが、それ以外のデータベースは語形の構成を問題とするので、語基を除いた部分（助詞・助動詞の連続）の異なりの一覧となる。従って、語形というよりは語型と呼ぶ方がふさわしいかもしれない。

３　動詞語形DBの作成

３．１　EDRコーパスの特徴

動詞語形データベースの作成にはEDRコーパス[3]を利用した。出典情報によると、EDRコーパスに採取された文の内訳は、新聞（約46%）・雑誌（約34%）・辞典（約11%）・用例集（約8%）である。各レコードは凡そ次のような構造を持つ[4]。

・EDRコーパス1.5版（文数：207,802文）
1. 文固有の情報（番号・出典など）
2. 文
3. 構成要素情報（以下の繰り返し）
   3.1. 表記
   3.2. かな表記
   3.3. 品詞
   3.4. 概念選択
4. 形態素情報
5. 構文情報
6. 意味情報
7. 管理情報

語形データベース作成には３の構成要素情報のうち表記と品詞の情報だけを利用した。

3．2　異なり動詞語形の抽出

次の手順により、まず動詞を先頭とする助詞・助動詞連続の異なり一覧（以下、異なり動詞語形と呼ぶ）を得た（図2参照）。

1. 構成要素情報の抽出　→　207,802 文
2. 動詞から始まる語形の抽出　→　597,823 語形
3. 接頭語のない動詞語形の抽出　→　597,239 語形
4. 品詞列・語形列の異なり抽出　→　3,609 語形

2は、(接頭語)動詞で始まり次の自立する形態素あるいは記号などまでを1レコードとするものである。次の形態素までを含めたのは、EDR コーパスには活用形や原形に関する情報がないことを補うためである。3で接頭語のない動詞語形のみを抽出したのは、接頭語の付いた語形は「お読みに（なる）」など、動詞というよりは名詞の語構成系列に属すると考えたからである。4の品詞列とは動詞（用言語基）部分を"動"に置き換え、次の語の直前までの形態素から接尾語を除いたもの、語形列とは表記に関して同様に処理したものである（下例参照）。

原文　：報道/さ/れ/な/かっ/た/の/だろ/う/か/と
語形列：動·れ·な·た·の·だろ·う·か·と
品詞列：動・助動詞・助動詞・助動詞・助詞・助動詞・助動詞・助詞・助詞

そして、この品詞列・語形列でソート・頻度付きの異なりを抽出して作成したのが異なり動詞語形一覧であり、次のような構造を持つ（例は次の形態素が出現せず文が終わる場合で、"・"で示してある）。

1. 頻度	217621
2. 代表レコードの番号	JCO005432
3. 文中の要素番号	10
4. 要素数（次品詞含む）	3
5. 語形（次品詞を含む）	ある
6. 次品詞名	・
7. 品詞列	動
8. 語形列	動

また、この他に原文を抽出したデータ（207,802文）、品詞列のみを抽出したデータ（同）、形態素別のデータ（5,105,047 形態素)も補助的に作成した。品詞別の形態素数は表1に示す通りである。

表1．品詞別　延べ語数

品詞	延べ語数	品詞	延べ語数
名詞	1,280,325	助詞	1,358,434
動詞	597,823	助動詞	305,562
形容詞	54,966	接頭語	21,554
形容動詞	58,286	接尾語	125,989
副詞	69,096	語尾	605,393
接続詞	22,418	数字	75,465
感動詞	314	記号	491,849
連体詞	37,573	合計	5,105,047

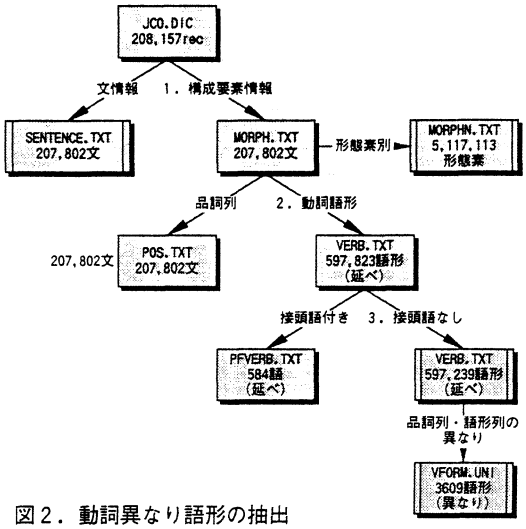


図2．動詞異なり語形の抽出

3．3　異なり動詞語形一覧の特徴

異なり語形一覧の、語尾を除く要素数別の語形数（表2）を見ると、要素数3～5の語形が圧倒的に多く約 87%を占めるが、延べ語形数では要素数1（動詞のみ）～3までで約 97%を占めている。割合簡単な語形が多いのは、扱っている文の大半が新聞・辞典などであることによるものかもしれない。

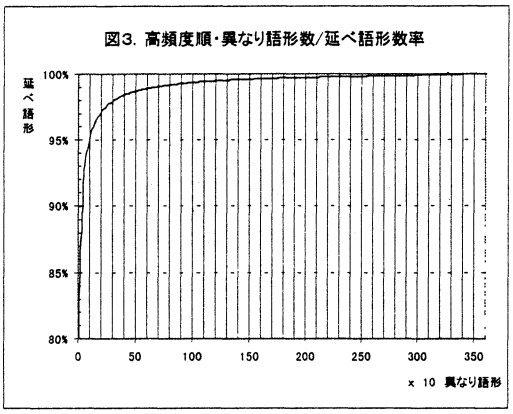
表2．要素数別　動詞語形数

要素数	異なり		延べ	
	語形数	割合	語形数	割合
1	1	0.03%	217,621	36.44%
2	245	6.79%	285,035	47.73%
3	1,098	30.42%	76,759	12.85%
4	1,305	36.16%	14,478	2.42%
5	723	20.03%	2,956	0.49%
6	206	5.71%	350	0.06%
7	26	0.72%	35	0.01%
8	3	0.08%	3	0.00%
9	2	0.06%	2	0.00%
合計	3,609	100.00%	597,239	100.00%

次に頻度別の語形数（次頁表3）を見ると異なり語形のほぼ半数が頻度1、つまり約 60 万語形中一度しか使われていない形であることが分かる。従って、これらの語形がカバーするのは全体の 0.3%にしか過ぎない。逆に言えば、高頻度・少数の語形が全体のほとんどをカバーしていることになる。次頁図3に示したように、高頻度順の異なり 10 語形で全体の 80%、40 語形で 90%、100 語形で 95%、650 語形で 99%をカバーする。ちなみに頻度 10,000 以上の語形は次の6語形である。

表３．頻度別　動詞語形数

頻度	動詞語形(異なり)				動詞語形(延べ)			
	語形数		累計		語形数		累計	
	異なり	割合	異なり	割合	延べ	割合	延べ	割合
1	1,804	49.99%	1,804	49.99%	1,804	0.30%	1,804	0.30%
～10	1,207	33.44%	3,011	83.43%	4,727	0.79%	6,531	1.09%
～100	442	12.25%	3,453	95.68%	14,299	2.39%	20,830	3.49%
～1,000	116	3.21%	3,569	98.89%	33,126	5.55%	53,956	24.79%
～10,000	34	0.94%	3,603	99.83%	94,123	15.76%	148,079	24.79%
～217,621	6	0.17%	3,609	100.00%	449,160	75.21%	597,239	100.00%
合計	3,609	100.00%			597,239	100.00%		



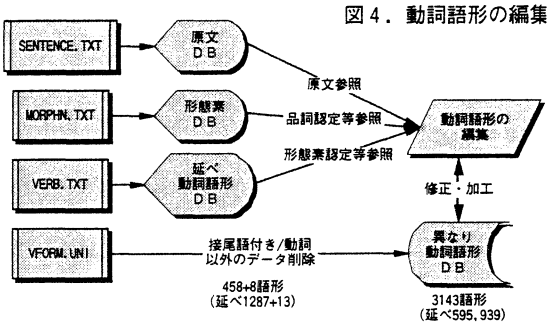
頻度	品詞・語形	例
217,621	動詞	ある
195,874	動詞・助詞「て」	よって
86,883	動詞・助動詞「た」	した
15,713	動詞・助詞「と」	捕獲すると
11,922	動詞・助動詞「れ」	2分され
11,687	動詞・助動詞「な」	あかない

３．４　異なり動詞語形一覧の問題点

前節で示した数値は EDR コーパスの情報だけから機械的に得られるものであるが、これには質の異なる二つの制約がある。第一は EDR コーパスの情報量による制約である。既に述べたように EDR コーパスには助詞・助動詞の細分についての情報や動詞・助動詞の活用形に関する情報がないため、(1)異なる語形が同じ語形と認定されてしまう場合がある。また、同様に原形に関する情報がないため、(2)同じ語（形態素）の別の語形（異形態）が異なる語形に認定されてしまう場合もある。これは言語研究のためのデータベースとしては満足できるものではない。

第二は、エラーに関する制約である。例えば少なめに見積もって構成要素情報に 0.01%の割合でエラーが含まれると仮定すると、動詞語形全体に含まれる形態素数は約 160 万形態素（1,588,373）であるため、約 160 のエラーが含まれる。エラーには(a)品詞認定の誤り、(b)語幹・語尾など単位切りの誤りの２種類があるが、どちらもエラー他と異なる語形列や品詞列として現れるため、延べ語形に対して与える影響は 0.02～0.03%であっても、異なり語形に対しては数%の影響を与えると考えられる。従って(3)低頻度の異なり動詞語形にエラーが多く含まれることになる。実際、頻度１の異なり動詞語形には数多くのエラーデータが含まれている。

図４．動詞語形の編集



３．５　動詞語形ＤＢの作成

そこで、これらの問題を解決するために人手による異なり動詞語形の修正を行い、動詞語形データベースを作成した（図４参照）。

まず、異なり動詞語形一覧から接尾語付きの動詞 458 語形を削除した。これは接頭語と同様、接続した接尾語により異なった系列の語構成をすると考えられるからであるが、同時に語尾を接尾語と誤るようなエラーデータを多く含むためでもある（品詞認定の判断に迷った際に、副詞や接尾語とすることもよく見られる）。また、明らかに動詞語形でないもの（品詞認定の誤り）８語形も削除した。

残りのデータに品詞列・語形列編集用のフィールドを追加、データベース化し、これに対して編集を行った。その際、図２に示した補助的なデータもデータベース化し、参照用に用いた。編集の主な点と例は次の通りである。

1. エラーの修正（品詞・単位切り）
2. 活用語に対する原形の付与（助動詞）
   たろ、た、たら→た：な→ない
   だろ、で～なら→だ：た→たい
3. 異形態の形態素への統合（助動詞）
   れ、られ→<ら>れる
   せ、させ→<さ>せる
   う、よう→<よ>う
4. 音便形接続の統合（助詞・助動詞）
   見て、読んで→て
   見た、読んだ→だ
5. ほぼ同じ形態素の統合
   けれど、けれども→けれど<も>
6. 同形の異なる形態素の区別
   そう→そうだ（様態）
   　→そうだ（伝聞）

実際の編集は、2～4の各項目ごとにまとめて行い、その際に発見したエラーを修正するという方法をとっている。さらに低頻度語形は集中的にエラーチェックを行った。この際、EDR コーパスの仕様に沿った形で編集を行ったが、品詞認定の範囲や基準が明示されていないために、学校文法[5][6][7]や辞典[8][9][10]を参考にしたり、延べ語形データ（当然、これらにもエラーが含まれる）を参照することにより EDR コーパスの仕様を仮定したが、細かい点は筆者の文法観が反映されてしまっているかもしれない。現在幾つかのチェックを行っているが、2173の異なり動詞語形を得ている。

3．6　動詞語形ＤＢの問題点

動詞語形データベースの編集に当たっては、作業効率の点から、出来る限り異なり語形ＤＢの範囲内で作業を行った。頻度1の異なり語形に関して問題はないが、頻度2以上の異なり語形を編集すると頻度情報が意味を持たなくなってしまう。

作業当初は延べ語形ＤＢあるいは形態素ＤＢまで遡って検討し頻度調整用のデータを作成していたのであるが、これでは約 60 万語形のすべてを編集することに等しい。また高頻度異なり語形について延べ語形の一部を調査したところ、問題を含むデータが予想以上に多いため方針を変更した。活用形情報についても同様の事情がある。これらは異なり語形のチェックを終了してから検討したいと考えている。

前節で述べた作業はデータに疑問があり、コーパスの仕様が明確でない以上、データのチェックと言うよりは文法を作る作業に近い。動詞語形データベースの一覧とともに編集作業方針（＝文法）の詳細は評価と併せて別の機会に発表したいが、編集を行った筆者の主観やエラーが紛れ込むことは避けがたい問題である。

4．今後の課題

以上、EDR コーパスを利用した動詞語形データベース作成の方法と問題点、動詞語形データベースの位置付けなどを報告した。電子化コーパスの利用には問題点も多いが、その制約を考慮しさえすれば大規模データを利用できる利点の方が大きい。今後は動詞語形のチェックを続け、異なり語形の一覧を確定するとともに、他のコーパス利用なども考慮しながら、他の活用語形データベースの作成を行っていく。また、語形データベースの評価及びこれを利用した文法研究を言語処理にフィードバックさせて行きたいと考えている。

付記. 本稿は 1999 年 1 月に上智大学国文学会[11]で発表した内容の一部を修正・発展させたものである。また、動詞語形データベース作成には熊本学園大学科学研究費[12]の助成を受けた。

参考文献

[1] 川田亮一、佐野洋：日本語語構成文法－形態論のモジュール化と標準化を目指して－、情報処理学会研究報告 Vol95, No.6（1995）
[2] 国立国語研究所：「現代雑誌九十種の用語用字(3)」所収、秀英出版（1964）
[3] 日本電子化辞書研究所：EDR 電子化辞書 1.5 版・日本語共起辞書付録
[4] 日本電子化辞書研究所：EDR 電子化辞書 1.5 版仕様説明書（1996）
[5] 橋本進吉：「改制　新文典別記　口語篇」冨山房（1938）
[6] 平岡敏夫他編：現代の国語Ⅱ、大修館書店（1998）
[7] 紅野俊郎他著：精選新国語Ⅱ現代文編、明治書院（1999）
[8] 西尾実他編：岩波国語辞典第5版、岩波書店（1994）
[9] 金田一春彦他編：学研国語大辞典、学習研究社（1978）
[10] 鈴木一彦他編：研究資料日本文法第7巻、助辞編(三)助詞・助動詞辞典、明治書院(1985)
[11] 川田亮一：日本語用言の活用体系－語構成論における動詞活用モデル－、平成十年度上智大学国文学会冬季大会(1999)
[12] 川田亮一、服部隆、馮蘊澤：自然言語処理を前提とした日本語活用体系についての形態論的基礎研究、平成九年度熊本学園大学科学研究課題（1997）